CLAIMS

What is claimed is:

1. A method for content mining of semi-structured documents comprising:

receiving a semi-structured document;

converting said semi-structured document to a document-type independent format;

analyzing formatting information of said semi-structured document;

adding information to said semi-structured document describing said semi-structured document's structure, based upon said analyzing; and

mining said semi-structured document for specified information, wherein said added information facilitates said content mining.

2. The method for content mining of semi-structured documents as recited in Claim 1, wherein said converting further comprises:

receiving said semi-structured document in a document-type dependent format; and

outputting said semi-structured document in a document-type independent format.

3. The method for content mining of semi-structured documents as recited in Claim 2, wherein said document-type independent format is the Extensible Markup Language (XML) format.

4. The method for content mining of semi-structured documents as recited in Claim 3, wherein said added information comprises an XML tag describing a feature of said semi-structured document's structure.

5. The method for content mining of semi-structured documents as recited in Claim 4, wherein said analyzing further comprises utilizing a plurality of said XML tags to derive said semi-structured document's structure.

6. The method for content mining of semi-structured documents as recited in Claim 5, wherein said mining comprises:

performing a query, wherein an extraction rule is provided defining a plurality of attributes of said specified information;

finding an XML tag which corresponds to at least one of said plurality of attributes; and

retrieving a value contained within said XML tag which corresponds to at least one of said plurality of attributes.

7. The method for content mining of semi-structured documents as recited in Claim 6 wherein said specified information comprises a plurality of said retrieved values.

8. A computer system comprising:

a bus;

a memory unit coupled to said bus; and

a processor coupled to said bus, said processor for executing a method for content mining of semi-structured documents, said method comprising:

receiving a semi-structured document;

converting said semi-structured document to a document-type independent format;

analyzing formatting information of said semi-structured document;

adding information to said semi-structured document describing said semi-structured document's structure, based upon said analyzing; and

mining said semi-structured document for specified information, wherein said added information facilitates said content mining.

9. The computer system as recited in Claim 8, wherein said deriving further comprises:

receiving said semi-structured document in a document-type dependent format; and outputting said semi-structured document in a document-type independent format.

10. The computer system as recited in Claim 9, wherein said document-type independent format is the Extensible Markup Language (XML) format.

11. The computer system as recited in Claim 10, wherein said added information comprises an XML tag describing a feature of said semi-structured document's structure.

12. The computer system as recited in Claim 11, wherein said analyzing further comprises utilizing a plurality of said XML tags to derive said semi-structured document's structure.

13. The computer system as recited in Claim 12, wherein said mining comprises;

performing a query, wherein an extraction rule is provided defining a plurality of attributes of said specified information;

finding an XML tag which corresponds to at least one of said plurality of attributes; and

retrieving a value contained within said XML tag which corresponds to at least one of said attributes.

14 The computer system as recited in Claim 13 wherein said specified information comprises a plurality of said retrieved values.

15. A computer-usable medium having computer-readable program code embodied therein for causing a computer system to perform a method for content mining of semi-structured documents comprising:

receiving a semi-structured document;

converting said semi-structured document to a document-type independent format;

analyzing formatting information of said semi-structured document;

adding information to said semi-structured document describing said semi-structured document's structure, based upon said analyzing; and

mining said semi-structured document for specified information, wherein said added information facilitates said content mining.

16. The computer-usable medium as recited in Claim 15, wherein said deriving further comprises:

receiving said semi-structured document in a document-type dependent format; and

outputting said semi-structured document in a document-type independent format.

17. The computer-usable medium as recited in Claim 16, wherein said document-type independent format is the Extensible Markup Language (XML) format.

18. The computer-usable medium as recited in Claim 17, wherein said added information comprises an XML tag describing a feature of said semi-structured document's structure.

19. The computer-usable medium as recited in Claim 18, wherein said analyzing further comprises utilizing a plurality of said XML tags to derive said semi-structured document's structure.

20. The computer-usable medium as recited in Claim 19, wherein said mining comprises;

performing a query, wherein an extraction rule is provided defining a plurality of attributes of said specified information;

finding an XML tag which corresponds to at least one of said plurality of attributes; and

retrieving a value contained within said XML tag which corresponds to at least one of said attributes.

21. The computer-usable medium as recited in Claim 20 wherein said specified information comprises a plurality of said retrieved values.